R을 이용한 전국 대학의 학과 명칭 분석

반재훈*, 하종수**

고신대학교 IT경영학과*, 경남정보대학 방송영상과**

Analysis of University Department Name using the R

ChaeHoon Ban*, JongSoo Ha**

Dept. of IT Management, Kosin University*

Department of Broadcasting & Imaging, Kyungnam College of Information & Technology**

E-mail : chban@kosin.ac.kr* , jsha@kit.ac.kr

요 약

스마트 정보 기기를 통해 사회 전 분야에서 대규모의 데이터가 생산되는데 이를 저장하고 분석하여 새로운 지식을 얻을 수 있는 빅데이터 처리기술은 사회의 여러 분야에서 중요성이 강조되고 있다. 이러한 빅데이터를 분석할 수 있는 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 R을 이용하여 전국에 2·4년제 대학, 대학원의 학과를 분석한다. 학과 명칭을 수집하고 각 데이터를 분석하여 학과 명칭의 빈도를 조사하며 대학에 어떤 학과 명칭이 자주 사용되는지를 파악한다.

키워드

Big Data, R, Text Mining, University Major, Analysis

Ⅰ. 서론

IT 기술의 발전에 따라 실생활에서 발생하는 대규모의 비정형 데이터를 수집하고 수집된 데이터를 이용하여 미래를 예측할 수 있는 빅데이터의 중요성이 강조되고 있으며, 다양한 산업에서 이를 활용하고 있다. 이러한 빅 데이터를 분석할 수 있는 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다.

대학은 최고의 학문기관으로서 시대의 발전과 요구에 따라 그에 대응하는 학과를 개설하고 유지해 왔다. 따라서 대학의 학과명을 분석하면 현 시대의 요구와 기술의 발전에 대하여 알 수 있다. 본 논문에서는 빅데이터 분석도구인 R을 이용하여 전국에 2·4년제 대학, 대학원의 학과를 분석한다. 학과 명칭을 수집하고 각 데이터를 분석하여 학과 명칭의 빈도를 조사하며 대학에 어떤 학과 명칭이 자주 사용되는지를 파악한다.

본 논문의 구성은 다음과 같다. 2장에서는 다양한 분야에서 빅데이터를 이용하여 문제를 해결한 관련연구를 기술한다. 3장에서는 본 논문에서 다루는 대학의 학과 정보를 R 프로그램을 활용하여 데이터를 분석하는 방법에 대해 기술한다. 4장에서는 각 대학의 학과의 빈도를 분석하고 이를 워드 클라우드 형태의 그래프로 표현하며, 마지막 5장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 관련연구

기존의 빅데이터 분석 기술로는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 소셜 마이닝 등 다양한 기법을 통한 빅 데이터 분석연구가 있었다. [1]에서는 정보통신의 발달과 소셜 미디어의 급속한 확산으로 생산된 빅 데이터를 분석하는 기법과 인프라 기술에 대해 기술하고 한글 텍스트 데이터를 R 프로그램을 이용하여 usesejongdic() 이라는 함수를 이용하여 명사만 추출하는 방법으로 비정형 데이터를 분석하였다. [2]에서는 데이터 시각화 도구 통계 패키지인 R을 이용하여 대기오염의 자료를 여러 가지 방법의 데이터 시각화를 통하여 나타내었고, 데이터 시각화 방법별로 통계적인 방법을 활용한 분석과 연계하여 어떤 특징이 있는지를 나타냈다. 2차원의 히스토그램과 선점도, 상자그림, 3차원 산점도와 투시도 등 다양한 방법의 그래프를 구현하여 오존농도와 설명 변수들 간에 어떠한 관련성이 있는지를 분석하였다.

[3]은 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특허 검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행했다. [4]에서는 성경의 텍스트 데이터를 성경전체, 구약성경, 신약성경, 모세오경, 사복음서 데이터 분석결과를 각각의 워드 클라우드 형태 그림으로 표현하여 성경데이터를 분석하여 성경을 읽는 독자에게 주는 메시지가 무엇인지에 대한 연구를 제시하였다.

III. 데이터 분석 방법

본 논문에서는 전국 대학의 학과를 빅데이터 분석도구인 R을 이용하여 워드 클라우드 형태의 그래프로 표현한다. 먼저 각 대학에 어떠한 학과가 있는지를 조사하기 위하여 대학 알리미(http://www.academyinfo. go.kr/)에서 제공하는 학과정보를 이용하였다. 이 데이터는 19열로 이루어진 45569개의 레코드로 구성된 데이터로 엑셀 형식이다.

원본 데이터는 전문대학, 대학, 대학원의 모든 학과로 구성되어 있는데 4년제 대학의 분석을 위하여 전문대학, 대학원, 폐과를 제외한 대학학과를 추출하였으며 레코드 개수는 9808개이다. 이 두 개의 데이터 셋을 가지고 1차 분석을 하였다. 데이터의 분석과정은 그림 1과 같다.

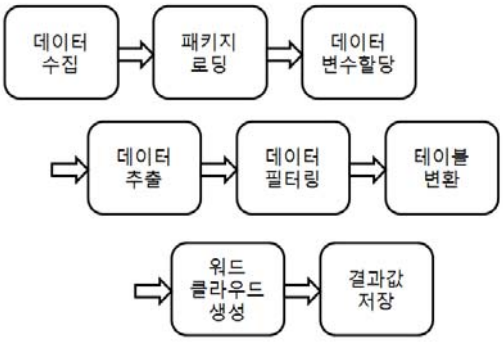

그림 1. 데이터 분석 과정

그러나 1차 분석 결과 같은 학과명임에도 불구하고 다르게 분석되는 학과들을 발견하였다. 예를 들어, 경영학과, 경영학부, 경영정보학과 등 같은 의미의 학과명임에도 불구하고 다른 학과(단어)로 인식되는 문제가 발생하여 보다 정확한 결과를 위하여 전처리 과정으로 필터링을 수행하였다. 필터링에서는 먼저 특수문자를 삭제(예 컴퓨터·전자에서 · 삭제)하였고 학과명에서 "학과", "학부"라는 어미를 삭제하였다.

그런데 어미를 삭제하는 경우에 수학과, 철학과 등의 학과명에서 학과를 삭제해서 의미가 회손되는 경우는 전처리를 하였으며 컴퓨터공학과, 컴퓨터과학과 등 공학, 과학 등의 단어로 끝나는 경우에도 학과를 삭제할 때 전처리하고 2차 분석을 실시하였다. 그러나 이러한 2차 분석에서도 단어를 완벽하게 분석하지 못하는 문제가 발견되었다. 예를 들어 "컴퓨터전자학과"의 경우 "컴퓨터"와 "전자"라는 두 개의 단어로 분리되어 분석되어야 하는데 이를 하나의 단어로 인식하는 한계가 발생하였다.

2차 분석에서의 문제를 해결하기 위하여 단어를 추출할 수 있는 패키지인 KoNLP를 이용하여 3차 분석을 실시하였다. 이때 사용한 사전은 useNIADic()로서 약 983012의 명사를 가지고 있는 사전이다. 이를 이용하여 분석한 결과 정확한 단어의 추출이 불가능하다는 것을 발견하였다. 예를 들어 기계항공정보융합학과의 경우에 "기계", "항공", "정보", "융합"이라는 단어가 추출되어야 하는데 "기계", "항공", "정보", "융", "합"으로 단어가 추출되는 문제가 발생하였다. 이는 사전을 어떠한 것을 사용하느냐에 따라 달라지므로 추후 연구에서는 사용자 사전을 구축하여 단어를 분석할 계획이며 본 논문에서는 1차와 2차 분석만을 기술한다.

IV. 데이터 분석 결과 및 비교

본 논문에서는 R을 이용하여 전국에 2·4년제 대학, 대학원의 학과를 분석하였다. 학과 명칭을 수집하고 각 데이터를 분석하여 학과 명칭의 빈도를 조사하였다.

학과명	빈도	학과명	빈도
사회복지학과	477	유아교육과	216
경영학과	447	건축공학과	209
간호학과	341	법학과	197
행정학과	308	기계공학과	192
컴퓨터공학과	229	영어영문학과	190

그림 2. 전국 전문대, 대학, 대학원 학과 분포

그림 2는 45569개의 전국 전문대, 대학, 대학원의 학과 분포를 워드클라우드 형태로 나타낸 분석 결과이다. 전처리를 거치지 않고 원본을 그대로 분석한 결과에서는 사회복지학과가 477회로 가장 많은 빈도를 보였으며 경영학과, 간호학과의 순서대로 많이 나타났다.

학과명	빈도	학과명	빈도
간호학과	128	식품영양학과	59
사회복지학과	114	영어영문학과	58
경영학과	85	컴퓨터공학과	58
유아교육과	66	경영학부	51
행정학과	62	국어국문학과	50

그림 3. 전국 대학 학과 분포

그림 3은 9808개의 전국 4년제 대학의 학과 분포를 워드클라우드 형태로 나타낸 분석 결과이다. 이 또한 전처리를 거치지 않고 원본을 그대로 분석한 결과이며 간호학과가 128회로 가장 많은 빈도를 보였으며 사회복지학과, 경영학과의 순서대로 많이 나타났다. 그러나 같은 의미의 학과임에도 다른 단어로 인식되어 경영학과와 경영학부가 다른 단어로 인식되는 문제점이 발견되었다.

학과명	빈도	학과명	빈도
경영	667	간호	364
사회복지	620	법	265
건축	533	기계	263
컴퓨터	448	정보통신	259
행정	384	전자	254

그림 4. 전처리후 전국 전문대, 대학, 대학원 학과 분포

그림 4는 전처리를 실시한 전국 전문대, 대학, 대학원의 학과 분포를 워드클라우드 형태로 나타낸 분석 결과이다. 경영이라는 단어가 667회로 가장 많이 나타나서 전처리를 실시하지 않은 경우의 사회복지학과의 477회와 다른 결과를 도출하였다. 이는 앞에서 언급한 바와 같이 경영학과, 경영학부 등을 같은 단어로 인식하여 분석하였기 때문이다.

그림 5는 전처리를 실시한 전국 4년제 학과 분포를 워드클라우드 형태로 나타낸 분석 결과이다. 경영이라는 단어가 162회로 가장 많이 나타나서 전처리를 실시하지 않은 경우의 간호학과의 128회와 다른 결과를 도출하였다.

학과명	빈도	학과명	빈도
경영	162	행정	80
사회복지	153	식품영양	71
건축	144	영어영문	71
간호	135	유아	66
컴퓨터	110	경제	65

그림 5. 전처리후 전국 대학 학과 분포

앞에서 언급한 바와 같이 이러한 전처리 후의 분석 방법도 의미있는 명사를 추출하여 처리하지 못한 한계점을 가지고 있다. 예를 들어 “컴퓨터전자학과”의 경우 “컴퓨터”와 “전자”라는 두 개의 단어로 분리되어 분석되어야 하는데 이를 하나의 단어로 인식하는 한계가 발생하였으며 이는 향후 연구에서 다룰 예정이다.

V. 결론 및 향후 연구

본 논문에서는 R을 이용하여 전국에 2·4년제 대학, 대학원의 학과를 분석하였다. 학과 명칭을 수집하고 각 데이터를 분석하여 학과 명칭의 빈도를 조사하며 대학에 어떤 학과 명칭이 자주 사용되는지를 파악하였다. 빅데이터 분석도구 R을 이용하여 수집된 데이터를 분석하고 이를 워드클라우드 형태의 그림으로 나타내어 시각화함으로써 빈도 수에 따른 키워드를 쉽게 알아 볼 수 있도록 하였다. 향후 연구에서는 학과의 단어들로 구성된 사용자 사전을 구축하고 학과명에 들어간 의미 있는 명사를 추출하여 새롭게 분석할 예정이다.

참고문헌

[1] 김현근, “R을 이용한 빅 데이터 사례 분석”, 호서대학교 일반대학원 정보통계학과 석사학위논문, 2014

[2] 오영창, 박은식, “R 소프트웨어를 이용한 대기오염 데이터의 시각화”, 한국데이터정보과학회지,vol. 26 no. 2, pp399-408, 2015

[3] 장청윤, 장정환, 김석주, 이현군, 이창호, “빅데이터 분석 도구 R을 활용한 효율적인 특허검색에 관한 연구”, 대한안전경영과학회지, vol. 15, no. 4, pp289-294, 2013

[4] 김용수, 반재훈 “성경 데이터를 활용한 빅데이터 분석”, 한국정보통신학회 2015 추계종합학술대회, pp349-352, 2015

[5] 반재훈, 이예찬, 안대중, 곽윤혁 “벤처창업 관련 뉴스 및 SNS 빅데이터 분석” 한국정보통신학회 2017 춘계종합학술대회, pp311-314, 2017